doi:10.3969/j.issn.1672-1292.2022.01.005

基于板块效应的深度学习股价走势预测方法

李庆涛,林培光,王基厚,周佳倩,张　燕,蹇木伟

(山东财经大学计算机科学与技术学院,山东 济南 250014)

[摘要]　股票价格预测作为金融预测领域中一项重要的研究方向,准确预测股票价格的涨跌可以帮助投资者盈利或及时止损. 经研究发现,某些因素(如政策、社会突发事件等)会对同板块下的多只股票价格产生影响,导致同板块的多只股票在某个时间段内出现相似的走势,即板块效应. 因此,同板块下多只股票的价格走势对于股票预测具有参考作用. 针对这一现象,提出了一种基于板块效应的深度学习股价走势预测方法. 首先,使用皮尔森(Pearson)相关系数和 XGBoost 算法对同板块下多只股票的收盘价进行分析,以筛选出与预测股票相关性高的多只股票,并使用自编码器对这些股票的收盘价进行降维,以提取股票的价格走势;其次,构建了一个基于卷积神经网络(convolutional neural networks,CNN)与长短期记忆(long short-term memory,LSTM)网络的混合深度学习预测模型,使用一维卷积神经网络提取输入数据的特征,使用 LSTM 网络对股票价格进行预测. 该模型使用银行、医药、酒业、娱乐传媒 4 个板块的股票作为实验数据集. 为了提高模型的预测效果,通过随机搜索对 LSTM 网络的神经元个数进行简单的分析,以选择较优的神经元个数. 最后,通过实验分析,基于同板块数据集的深度学习预测模型具有良好的预测效果.

[关键词]　同板块股票特征,XGBoost,股票预测,LSTM,深度学习
[中图分类号]TP391　[文献标志码]A　[文章编号]1672-1292(2022)01-0030-09

Deep Learning Stock Price Forecasting Method Based on Plate Effect

Li Qingtao,Lin Peiguang,Wang Jihou,Zhou Jiaqian,Zhang Yan,Jian Muwei

(School of Computer Science and Technology,Shandong University of Finance and Economics,Jinan 250014,China)

Abstract:As an important research direction in the field of financial forecasting,accurate prediction of stock price rise and fall can help investors to make profits or stop losses in time. It has been found that certain factors(such as policies,social emergencies)can have an impact on the prices of multiple stocks in the same sector,resulting in similar movements of multiple stocks in the same sector in a certain period of time,i.e. the sector effect. Therefore,the price trends of multiple stocks under the same segment are useful for stock forecasting. To address this phenomenon,a deep learning stock price trend prediction method based on the plate effect is proposed. Firstly,the Pearson correlation coefficient and XGBoost algorithm are used to analyze the closing prices of many stocks in the same sector so as to screen out the stocks with high correlation with the predicted stocks. Then,the autoencoder is used to reduce the dimension of the closing prices of these stocks,so as to extract the price trend of the stocks. Secondly,a hybrid deep learning prediction model based on convolutional neural network and long short-term memory network is constructed. One-dimensional convolutional neural network is used to extract the features of input data,and LSTM network is used to predict stock prices. The model uses stocks in four sectors,namely, banking, pharmaceuticals, alcohol, and entertainment media, as the experimental data set. In order to improve the prediction effect of the model,the number of neurons of the LSTM network is simply analyzed by random search to select the better number of neurons. Finally,the experimental analysis shows that the deep learning prediction model based on the same board dataset has good prediction effect.

Key words:characteristics of the same industry stocks,XGBoost,stock prediction,LSTM,deep learning

股票作为人们日常生活投资中重要的部分,一直备受投资者的关注. 由于股票价格具有高度的不稳定性以及复杂性,准确预测股票的价格走势仍然是一项艰巨的任务. 在金融预测研究早期,投资者与研究人员通过分析股票自身的收盘价、开盘价、最高价、最低价等特征,使用统计学习方法对于股票价格进行预

收稿日期:2021-08-31.
基金项目:国家自然科学基金项目(61802230).
通讯作者:林培光,博士,副教授,研究方向:信息检索、自然语言处理、机器学习. E-mail:llpwgh@163.com

测,并取得了一定的成果[1]. 但考虑到股票价格具有多特征以及非线性等特点,需要对股票数据进行进一步分析,以提取其中蕴含的隐藏信息. 例如:Zhou 等[2]通过在数据集中引入随机震荡指数、异同移动平均线等常用的股票技术指标,并使用 BP 神经网络对股票的收盘价进行预测,实验结果表明,引入股票的技术指标等特征可以提高模型的预测精度.

伴随着机器学习以及深度学习的快速发展,机器学习方法以及人工神经网络开始应用在金融预测领域[3]. 相较于传统的金融预测方法,神经网络在预测过程中可以更好地处理大量、高维的数据还能考虑到数据的时序特征以及非线性的特点,在金融预测领域得到了广泛的应用[4].

然而在股票市场中,股票并非独立存在,市场内的股票相互影响,存在板块效应的现象. 经过研究发现,同板块的多只股票在受到政策以及某些市场因素的影响下会出现明显的相关性趋势[5]. 考虑到上述现象,本文提出了一种基于板块效应的深度学习股价走势预测模型,选取银行、医药生物、酒业、娱乐传媒 4 个板块的 10 只股票作为实验数据集,分别选择中国银行(601988)、光大银行(601818)、新华传媒(600825)、华数传媒(000156)、上海医药(601607)、青岛啤酒(600600)作为预测股票,剩余股票作为同板块股票,通过 Pearson 相关系数、XGBoost 以及自编码器对输入的多只股票进行筛选降维,使用 CNN-LSTM 网络对股票价格进行预测. 实验结果表明,本文提出的模型优于传统的金融预测模型.

本文的主要贡献包括 2 个方面:

(1)不同于以往使用单只股票进行分析,考虑到同板块的板块效应以及相关性趋势,本文使用同板块下的多只股票作为输入数据,以提高模型的预测效果.

(2)本文提出了一种基于板块效应的深度学习股价走势预测模型,实验结果表明本文提出的模型相较于传统的金融预测模型具有更好的预测效果.

1 相关工作

传统股票预测方法使用的股票数据只有收盘价、开盘价、最高价、最低价、交易量等少量的特征,如何从少量的特征中获取更多有用的信息一直是投资者和金融研究人员关注的话题. Huang 等[6]从股票的基本财务比率入手,使用前馈神经网络(feedforward neural network,FNN)以及自适应神经模糊推理系统(adaptive network-based fuzzy inference system,ANFIS)构建了一个股票投资组合模型. 实验结果表明,FNN 和 ANFIS 所选择投资组合可以在金融市场中获得一定的收益. Li 等[7]以香港股票为例,在数据集中引入了新闻情感分析,并建立了 4 种情绪词典,实验结果表明文中构建的金融情绪词典可以大幅度提高模型的预测性能. Yu 等[8]为了进一步提取股票信息并减少输入数据冗余,通过使用主成分分析方法(principal component analysis,PCA)对股票数据进行降维处理,取得了较好的成果. 但 PCA 属于线性降维方法,并不能很好提取股票价格中的非线性关系,因此,Bao 等[9]提出了一个基于堆栈自编码器的深度学习预测模型,通过在数据预处理阶段使用堆栈自编码器对股票数据进行降维,通过 LSTM 网络对股票价格预测. 结果表明,该模型在预测精度和盈利能力方面均优于其他对比模型. XGBoost[10]、决策树等传统的机器学习方法也广泛应用在金融序列的特征提取方向.

在金融预测领域,金融预测方法主要分为传统的时序建模方法、机器学习方法、深度学习方法. 其中,传统的时序建模方法主要为自回归模型(auto regressive model,AR)、差分自回归移动平均模型[11](auto regressive integrate moving average model,ARIMA)等. 传统的时序建模方法主要是对线性数据进行预测,对于股票数据并不能有很好的预测效果. 机器学习方法主要为贝叶斯网络、支持向量机(support vector machine,SVM)[12]等,Xiao 等[13]通过使用奇异谱分析以及 SVM 对 2009 年到 2012 年上证指数的收盘价进行预测. 深度学习方法主要为 CNN[14]、RNN[15]、LSTM[16-17]等神经网络. 在时序预测领域,LSTM 网络通过循环机制以体现数据的时序关系,并使用细胞状态和三道“门”解决了 RNN 网络中梯度消失等现象. Ding 等[18]通过使用 LSTM 对股票的收盘价、最低价、最高价进行同时预测,实验结果显示,模型的预测精度在 95%以上. 由于单神经网络容易出现过拟合或欠拟合现象,为了提高模型的预测效果,Skehin 等[19]提出了 ARIMA 加 LSTM 结合的预测方法,通过使用 ARIMA 过滤股票价格中的线性趋势,再通过 LSTM 对过滤后结果进行进一步的预测,实验结果表明,该方法相对于传统的金融预测方法,该模型可以进一步提高金融数据的预测准确度. Mehtab 等[20]则提出了一种使用 CNN 与 LSTM 两种神经网络相结合的股票预测模型,

进一步证明了深度学习模型融合可以进一步提高预测效果.

2　模型结构

针对同板块下的多只股票,本文通过 XGBoost 算法与 Pearson 相关系数对输入数据进行筛选,筛选结果由 XGBoost 特征重要程度排序以及相关系数大小决定. 然后通过去噪自编码器模型对筛选后的数据进行降维处理,以提取同板块下多只股票的趋势. 最后,通过 CNN-LSTM 网络对股票的收盘价进行预测,为了提高预测精度,本文采用了随机搜索的方式对 LSTM 网络中的神经元参数进行分析,以提高模型的预测精度. 本文的模型结构如图 1 所示.

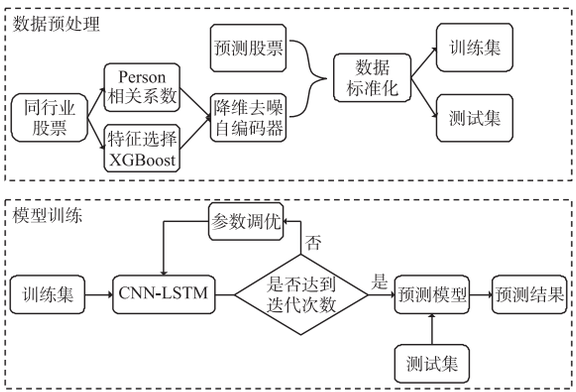


图 1　股票预测模型

Fig. 1　The stock forecasting model

2.1　同板块股票相关性筛选

为了分析板块内股票之间的相关性联系,本文使用了 XGBoost 以及 Pearson 相关系数对预测股票与同板块股票之间的相关性进行分析筛选.

根据股票市场的板块效应,板块内的个股会发生同步波动的现象,但并不是同一板块中的所有股票都会发生联动效应,同一板块下的多只股票存在相关性联系,板块内股票的相关性越强,则股票之间的联动性越强. 反之,股票之间的联动性越弱.

2.1.1　Pearson 相关系数

Pearson 相关系数是最为常用的相关系数之一,在股票领域也得到了广泛的应用. Pearson 相关系数的计算过程如式(1)所示,X,Y 分别为预测股票的收盘价以及同一板块内其他 9 只股票的收盘价,E 代表数学期望,cov 代表协方差. 当相关系数为正时,代表预测股票与同一板块内的股票存在正相关. 反之,则存在负相关. 当相关系数的绝对值越接近于 1,则代表股票之间的相关性越强. 一般来说,相关系数在 0.8~1.0 代表极强相关;0.6~0.8 为强相关;0.4~0.6 为中等程度相关;0.2~0.4 弱相关;0.0~0.2 极弱相关或无相关. 在本文中选择保留与预测股票相关系数在 0.8 的同板块股票.

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}. \tag{1}$$

2.1.2　XGBoost

Pearson 相关系数主要是分析数据之间的线性关系,为了进一步分析同板块下各股票对于预测股票的影响程度. 本文使用 XGBoost 算法对各板块下的股票的收盘价进行重要性排序.

XGBoost 算法是由 Chen 等[21]在 2016 年提出的一种集成学习算法. XGBoost 算法是在梯度提升决策树(gradient boosting decision tree,DBGT)上进行改进的. 对于 DBGT 模型,其做法是不断地拟合残差,直到误差的损失在可接受范围内.

在梯度提升树被创建后,可以相对直接地得到每个属性的重要性得分. 在模型训练的过程中,每棵树可以根据评估指标对输入的同板块的 9 只股票的收盘价进行打分,最后将每棵树对应的得分相加,即可得到重要性排序.

训练集 $D=\{(X_1,y_1),(X_2,y_2),\cdots,(X_N,y_N)\}$,$X$ 代表同板块 9 只股票的收盘价特征,y 为预测股票的收盘价. XGBoost 模型提供了 3 种特征重要性的计算方法:Weight,Gain,Coverage. Weight 是特征在所有树中作为划分属性的次数;Gain 是使用特征在作为划分属性时 loss 平均的降低量;Coverage 是使用特征在作为划分属性时对样本的覆盖度. 本文选择使用 Gain 作为特征重要性计算方法. 其中 Gain 计算方法如式(2)所示. G_j 代表损失函数一阶导的总和,H_j 为损失函数二阶导的总和,λ 和 γ 为自定义常数. 在式(2)中,第 1 项为划分后左叶子结点的分值,第 2 项为划分后右叶子结点的分值,第 3 项为划分前该结点的分值,γ 为将该结点划分为叶子结点的复杂度.

$$\text{Gain}=\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{G_L^2+G_R^2}{H_L+H_R+\lambda}-\gamma. \tag{2}$$

2.2 数据降维

为了进一步提取同板块股票之间的趋势性,减少运算时间,使用去噪自编码器对经过 XGBoost 和 Pearson 相关系数筛选后的同板块股票的收盘价进行降维. 自编码器由于引入了神经网络和激活函数,可以实现对数据的非线性降维,在各个领域得到了广泛的应用. 去噪自编码器相较于传统的自编码器,在输入的时候给数据增加了一个噪声,以提高模型的鲁棒性. 因此,本文采用去噪自编码器对数据进行降维. 图 2 为传统的去噪自编码器的模型结构图.

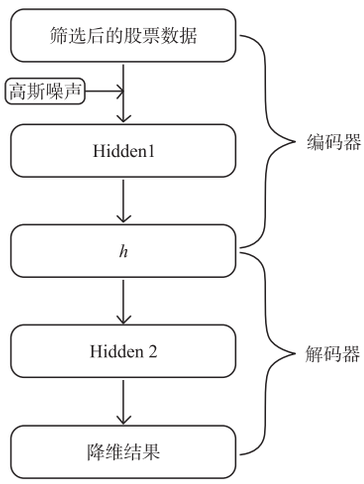


图 2 去噪自编码器模型

Fig. 2 Denoising autoencoder model

去噪自编码器通过编码器把筛选后的同板块股票映射到输出表征 h,在通过解码器将输出表征 h 映射到 $\hat{x}$. 其中 $\hat{x}$ 为重构数据,h 为降维之后的结果. 在模型训练过程中,本文在编码器与解码器中使用了 3 层全连接网络,以提高模型的学习能力,使用均方误差作为损失函数,Adam 作为优化器. 为了更好地可视化降维后的结果,本文将输出表征 h 设置为 1. 即将筛选后的股票特征降维到一维.

2.3 CNN-LSTM 模型

由于卷积神经网络在特征提取中具有良好的效果,长短期记忆网络在时序预测中表现较好. 本文选择使用一维卷积神经网络与长短期记忆网络的混合模型作为预测模型. 对于输入的股票数据,本文首先使用一维卷积神经网络对输入的数据通过卷积的方式进行局部的特征提取,以提高模型的预测效果,再将提取之后的数据输入到 LSTM 网络中以进行预测,最后通过全连接层输出. 并设置 Dropout 层以防止过拟合现象.

对于 CNN-LSTM 模型,其输入的数据由两部分拼接组成,第 1 部分为经过筛选降维之后的同板块股票的收盘价,第 2 部分为预测股票自身的最高价、最低价、开盘价、收盘价. 本文在预测过程中使用双层的 LSTM 网络,以提高模型的预测效果,加入了 Dropout 层以防止模型出现过拟合现象. LSTM 网络的公式如下所示. 其中 x_t 为了经过一维卷积网络特征提取后的数据,$\boldsymbol{h}$ 为输出向量,f、i、o 分别为遗忘门、输入门、输出门,c_t 为细胞状态,σ 为 sigmoid 激活函数,$\boldsymbol{W}$ 代表权重矩阵,b 代表偏置.

$$f_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},x_t]+b_f), \tag{3}$$

$$i_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},x_t]+b_i), \tag{4}$$

$$\bar{c}_t=\tan h(\boldsymbol{W}_c\cdot[\boldsymbol{h}_{t-1},x_t]+b_c), \tag{5}$$

$$c_t=f_t*c_t-1+i_t*\bar{c}_t, \tag{6}$$

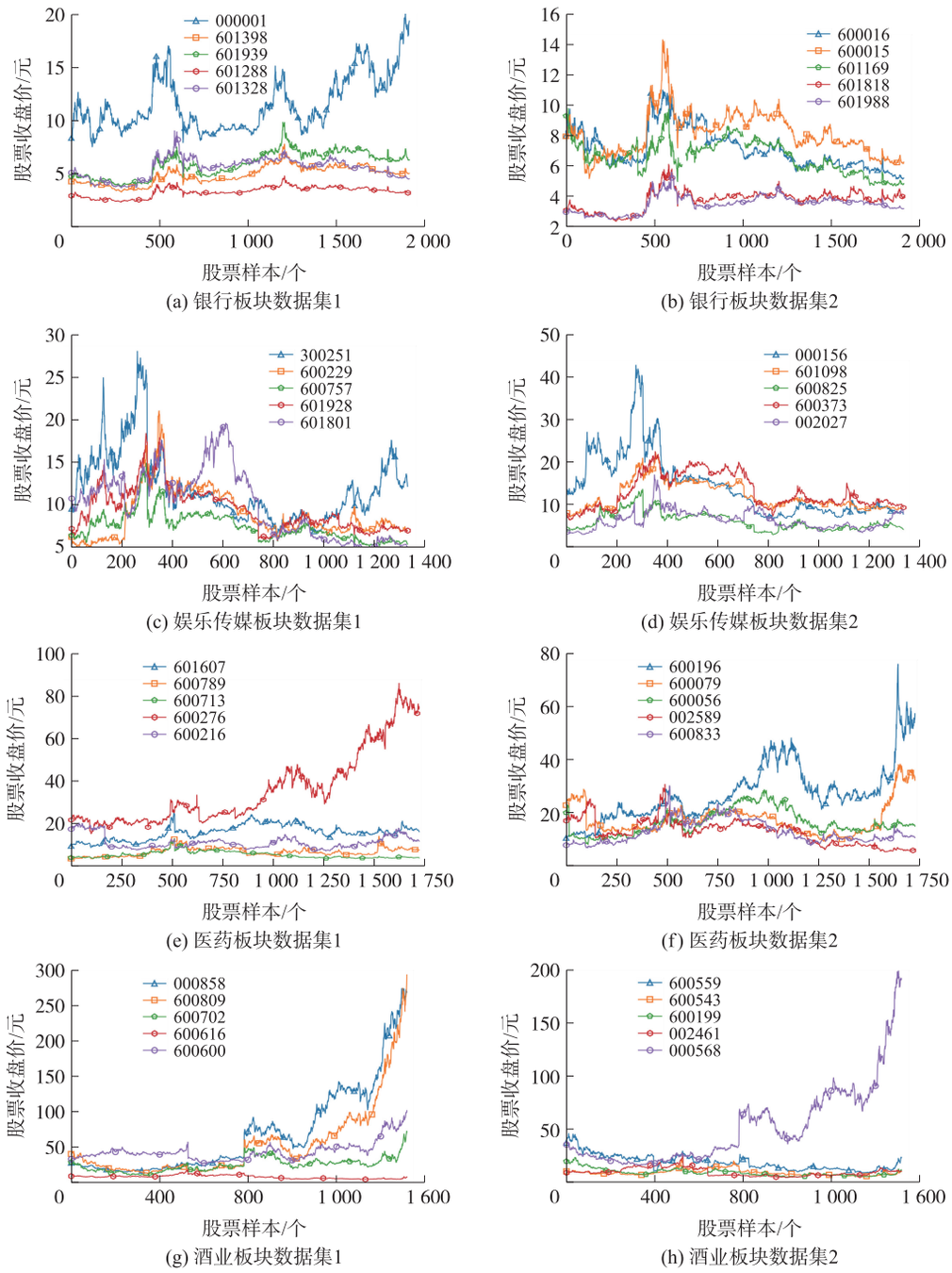
$$O_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},x_t]+b_o), \tag{7}$$

$$\boldsymbol{h}_t=o_t*\tan\boldsymbol{h}(c_t). \tag{8}$$

3 数据与实验分析

3.1 数据分析

本节从银行、酒业、医药、娱乐传媒 4 个板块中共选取 40 支股票,选取 2013 年 1 月 1 日到 2020 年 12 月 31 日,一共 7 年的股票日数据作为实验数据集票. 在每个板块中,每次选取一只股票作为预测股票,其余股票作为同板块股票. 为了增强实验的有效性,银行板块分别选取中国银行、光大银行作为预测股票,娱乐传媒板块分别选取新华传媒、华数传媒作为预测股票,酒业板块选择青岛啤酒作为预测股票,医药板块选择上海医药作为预测股票. 选取数据集前 80%作为训练集,后 20%作为测试集. 本文使用前 5 天的数据预测下一天. 图 3 为各板块股票收盘价价格走势. 从图 3 可以看出,同板块的股票在某个时间点会出现多数股票同时上涨或下跌趋势. 例如银行板块在横坐标 1200 时大多数股票出现了上升的趋势,酒类板块在横坐标 900 到 1 000 左右大多数股票都出现了下降的趋势. 为了进一步提取股票之间的相关性特征,本文对输入的股票进行了筛选,筛选的结果为 Pearson 相关系数大于 0.8,XGBoost 特征重要度排序累计占比 90%以上的特征. 最后将筛选后的股票输入到去噪自编码器中,以提取股票之间的趋势.



(a) 银行板块数据集1　(b) 银行板块数据集2

(c) 娱乐传媒板块数据集1　(d) 娱乐传媒板块数据集2

(e) 医药板块数据集1　(f) 医药板块数据集2

(g) 酒业板块数据集1　(h) 酒业板块数据集2

图 3　各板块股票数据集走势图

Fig. 3　Stock data sets of various plates chart

本文使用 XGBoost 算法对各板块下的股票进行重要性排序. 结果如图 4 所示.

图 4 的纵坐标为同板块下的 9 只股票与预测股票的特征重要程度排序结果,横坐标为输入的特征,横坐标的特征顺序与相关系数矩阵一致. 为了尽可能提取多只股票的趋势,本文选择特征重要程度累计超过 90%,作为筛选条件.

各板块相关系数矩阵如图 5 所示,从图 5 可以看出,通过相关系数矩阵,同板块各股票之间存在一定的相关性联系,部分股票之间出现强相关系数.

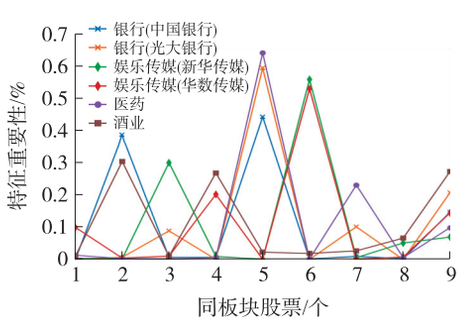


图 4　XGBoost 特征重要程度排序

Fig. 4　Importance ranking of XGBoost features

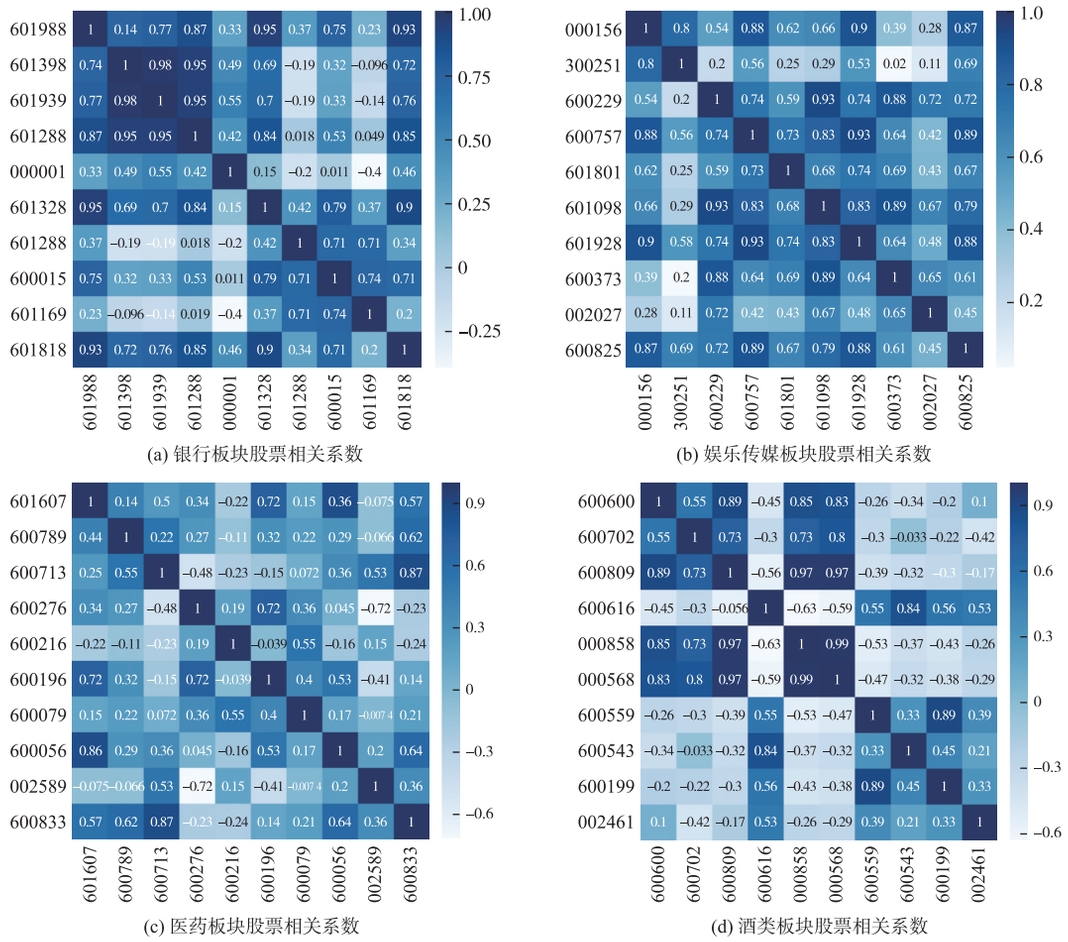


图 5　各板块股票数据集相关系数图

Fig. 5　The correlation coefficient graph of the stock dataset of each plate

3.2　评价指标与超参数分析

为了验证本文提出模型的可行性,本文选择银行、酒业、医药、娱乐传媒 4 个板块的股票作为实验数据集,以平均绝对误差(mean absolute error,MAE)、均方误差(mean square error,MSE)、均方根误差(root mean squared error,RMSE)、决定系数(R^2)作为模型的评价指标.

3.2.1　模型参数指标

本文选择使用 MAE、MSE、RMSE、R^2 作为评价指标. 以下为各评价指标的公式,

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i|, \tag{9}$$

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2, \tag{10}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}, \tag{11}$$

$$R^2=1-\frac{\sum_i(\hat{y}^{(i)}-y_i)^2}{\sum_i(\bar{y}-y_i)^2}. \tag{12}$$

式中,$\hat{y}_i$ 代表预测数据,y_i 代表真实数据,$\bar{y}$ 代表均值. 在评价指标中,MAE、MSE、RMSE 评价指标值越低,代表模型的拟合效果更好;R^2 评价指标越高,代表模型效果越好.

3.2.2　LSTM 网络神经元参数分析

为了提高模型的预测效果,本文使用随机搜索以选择较好的 LSTM 网络的神经元个数. 考虑到时间和运算性能. LSTM 网络神经元个数的搜索范围选择在 16−128 之间. 表 1 为各板块训练模型最优的 LSTM

神经元个数.

表 1　LSTM 神经元个数

Table 1　Number of LSTM neurons

种类	银行板块(中国银行)	银行板块(光大银行)	娱乐传媒板块(新华传媒)	娱乐传媒板块(华数传媒)	医药板块	酒业板块
LSTM_1	25	69	25	114	53	61
LSTM_2	114	68	114	102	60	123

3.3　实验结果分析

为了验证本文提出的模型的可行性,以中国银行、光大银行、新华传媒、华数传媒、上海医药、青岛啤酒 6 只股票作为实验数据集,并选择目前常用的金融预测神经网络 PCA-LSTM、Attention-LSTM、LSTM、CNN、多层感知机(multilayer perceptron,MLP)作为实验对比模型,各模型的预测指标结果如表 2 所示.

从表 2 可以得出,本文所提出的模型在各项预测指标中都要好于传统的金融序列预测模型,在预测指标的对比中,本文所提出的模型在 MAE、MSE、RMSE 指标均低于对比模型,R^2 指标均高于对比模型,证明了相较于传统的金融预测模型,本文所提出的模型具有更好的预测效果. 图 6 为数据集的部分实验结果.

表 2　实验结果表

Table 2　Table of experimental results

		MAE	MSE	RMSE	R^2
银行板块(中国银行)	Our model	0.020 6	0.001 0	0.032 2	0.966 0
	PCA-LSTM	0.022 6	0.001 3	0.036 6	0.958 2
	Attention-LSTM	0.022 7	0.001 2	0.034 4	0.958 9
	LSTM	0.022 7	0.001 4	0.036 9	0.956 3
	CNN	0.023 4	0.001 4	0.037 8	0.951 6
	MLP	0.033 0	0.002 0	0.044 6	0.936 5
银行板块(光大银行)	Our model	0.072 2	0.009 0	0.094 8	0.864 2
	PCA-LSTM	0.088 7	0.014 4	0.119 8	0.754 2
	Attention-LSTM	0.089 3	0.010 9	0.104 5	0.846 8
	LSTM	0.092 5	0.013 9	0.117 8	0.778 8
	CNN	0.136 7	0.026 2	0.161 9	0.546 3
	MLP	0.106 9	0.016 4	0.128 1	0.741 9
娱乐传媒板块(新华传媒)	Our model	0.120 1	0.029 2	0.170 9	0.903 2
	PCA-LSTM	0.150 7	0.042 8	0.206 8	0.853 9
	Attention-LSTM	0.132 1	0.034 9	0.186 9	0.885 8
	LSTM	0.151 8	0.042 5	0.206 2	0.852 6
	CNN	0.153 9	0.045 9	0.214 2	0.845 7
	MLP	0.143 5	0.037 1	0.192 7	0.874 7
娱乐传媒板块(华数传媒)	Our model	0.477 7	0.436 9	0.661 0	0.987 2
	PCA-LSTM	0.509 5	0.580 6	0.762 0	0.984 7
	Attention-LSTM	0.523 2	0.580 0	0.761 6	0.983 6
	LSTM	0.647 7	0.952 2	0.975 8	0.974 5
	CNN	1.034 8	2.050 3	1.431 9	0.937 7
	MLP	0.752 2	1.217 4	1.103 4	0.964 4
医药板块(上海医药)	Our model	0.270 6	0.156 2	0.395 2	0.911 1
	PCA-LSTM	0.301 2	0.178 5	0.422 5	0.889 8
	Attention-LSTM	0.301 3	0.185 3	0.430 5	0.891 3
	LSTM	0.307 2	0.176 3	0.419 9	0.900 0
	CNN	0.349 9	0.249 4	0.499 4	0.850 5
	MLP	0.350 0	0.240 7	0.490 6	0.855 9
酒业板块(青岛啤酒)	Our model	0.912 3	1.410 0	1.187 4	0.957 7
	PCA-LSTM	0.913 7	1.549 2	1.244 7	0.946 8
	Attention-LSTM	1.020 4	1.842 8	1.357 5	0.937 5
	LSTM	1.046 7	2.010 3	1.417 9	0.934 5
	CNN	1.020 9	1.988 3	1.410 1	0.929 3
	MLP	1.153 0	2.313 3	1.521 0	0.927 6

如图 6 所示,虚线为对比模型的预测结果,实线为股票真实价格和本文所提出的模型的预测结果. 从表 2 以及图 6 可以看出,在金融预测领域,相对于 CNN 模型,LSTM 模型可以取得更好的效果,相较于传统的金融预测模型,本文提出的模型具有更好的拟合能力和较优的预测指标.

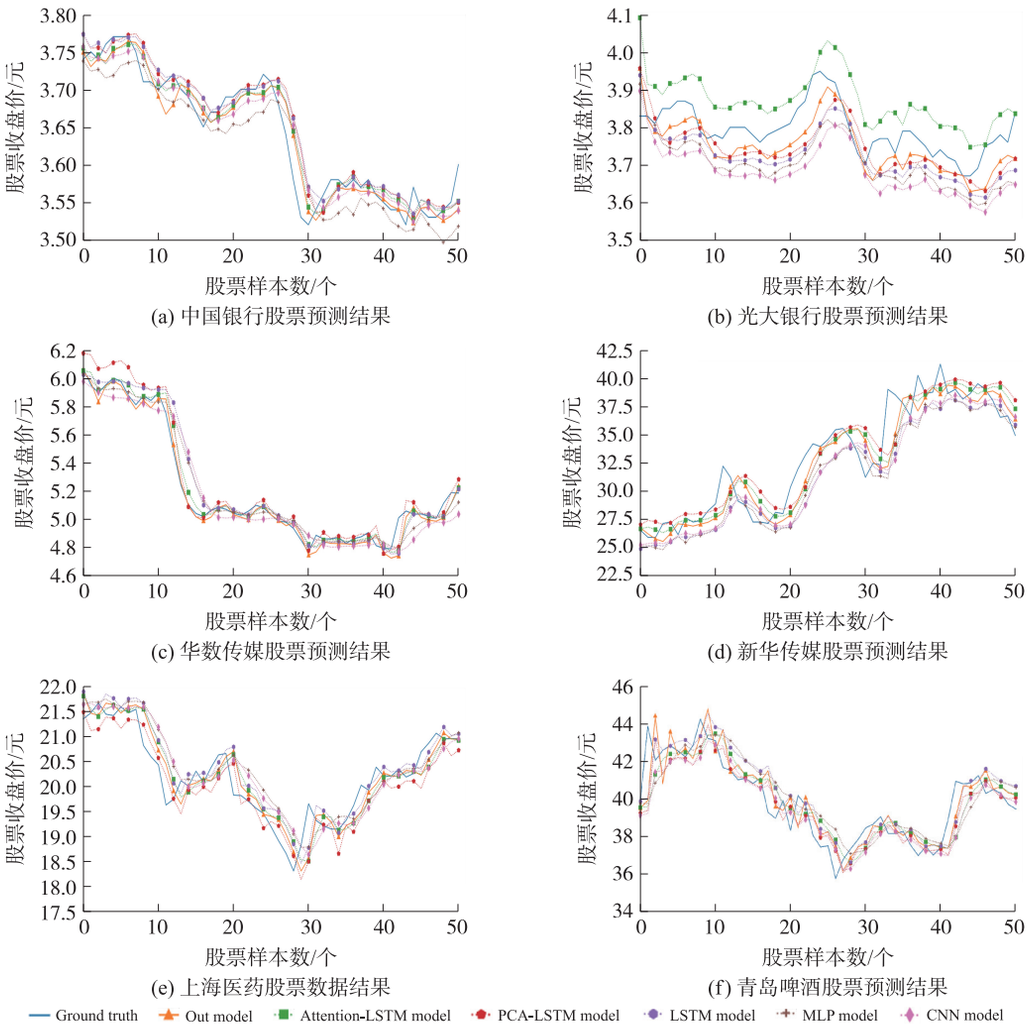


图 6　各模型实验结果图(部分)

Fig. 6　Diagram of experimental results of each model(partial)

4　结论

本文从同板块下多只股票之间存在相关性的角度出发,提出了一种基于板块效应的深度学习股价走势预测方法. 该模型通过使用 XGBoost 和 Pearson 相关系数对同板块下的多只股票进行筛选,使用去噪自编码器对筛选后的股票进行降维,最后通过使用 CNN 和 LSTM 网络对股票价格进行预测. 通过与目前常用的金融序列预测模型对比,本文提出的模型在预测指标以及拟合程度上均好于对比模型. 未来本文将不局限于对 LSTM 网络参数的探讨,并在数据集中引入情感、技术指标等股票的常用特征.

[参考文献](References)

[1] WEI D. Prediction of stock price based on LSTM neural network[C]//Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing(AIAM). Dublin:IEEE,2019:544-547.

[2] ZHOU Y X,ZHANG J. Stock data analysis based on BP neural network[C]//Proceedings of the 2010 Second International Conference on Communication Software and Networks,Singapore:IEEE,2010:396-399.

[3] 刘博,王明烁,李永,等. 深度学习在时空序列预测中的应用综述[J]. 北京工业大学学报,2021,47(8):925-941.

[4] 李金轩,杜军平,薛哲. 基于多视角股票特征的股票预测研究[J]. 南京大学学报(自然科学),2021,57(1):68-74.
[5] 张成云,汪俊. 北京奥运会对 A 股奥运板块的上市公司效益影响的研究[J]. 特区经济,2007(9):109-111.
[6] HUANG Y, CAPRETZ L F, HO D. Neural network models for stock selection based on fundamental analysis[C]// Proceedings of the 2019 IEEE Canadian Conference of Electrical and Computer Engineering(CCECE), Edmonton, Canada: IEEE,2019:1-4.
[7] LI X,WU P,WANG W. Incorporating stock prices and news sentiments for stock market prediction:a case of Hong Kong[J]. Information Processing & Management,2020,57(5):102212.
[8] YU H,CHEN R,ZHANG G. A SVM stock selection model within PCA[J]. Procedia Computer Science,2014,31:406-412.
[9] BAO W,YUE J,RAO Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory[J]. PloS One,2017,12(7):e0180944.
[10] CHEN C,ZHANG Q,YU B,et al. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier[J]. Computers in Biology and Medicine,2020,123:103899.
[11] 李冰. 沪深 300 股指预测——基于 ARIMA 模型和人工神经网络模型相结合的方法[D]. 广州:暨南大学,2018.
[12] 李辉,赵玉涵. 基于 DFS-BPSO-SVM 的股票趋势预测方法[J]. 软件导刊,2017,16(12):147-151.
[13] XIAO J,ZHU X,HUANG C,et al. A new approach for stock price analysis and prediction based on SSA and SVM[J]. International Journal of Information Technology and Decision Making,2019,18(1):287-310.
[14] 徐月梅,王子厚,吴子歆. 一种基于 CNN-BiLSTM 多特征融合的股票走势预测模型[J]. 数据分析与知识发现,2021,5(7):126-137.
[15] 陈佳. RNN 神经网络在股指预测中的应用研究[D]. 天津:天津科技大学,2019.
[16] 林培光,周佳倩,温玉莲. SCONV:一种基于情感分析的金融市场趋势预测方法[J]. 计算机研究与发展,2020,57(8):1769-1778.
[17] 胡聿文. 基于优化 LSTM 模型的股票预测[J]. 计算机科学,2021,48(Suppl 1):151-157.
[18] DING G,QIN L. Study on the prediction of stock price based on the associated network model of LSTM[J]. International Journal of Machine Learning and Cybernetics,2020,11(6):1307-1317.
[19] SKEHIN T, CRANE M, BEZBRADICA M. Day ahead forecasting of FAANG stocks using ARIMA, LSTM networks and wavelets[C]//Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science. Dublin:CEUR Workshop Proceedings,2018.
[20] MEHTAB S,SEN J. Stock price prediction using CNN and LSTM-based deep learning models[C]//Proceedings of the 2020 International Conference on Decision Aid Sciences and Application(DASA). Sakheer,Bahrain:IEEE,2020:447-453.
[21] CHEN T Q,GUESTRIN C. Xgboost:a scalable tree boosting system[C]//Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. New York,USA:Association for Computing Machinery,2016:785-794.

[责任编辑:陈　庆]